ESTIMATION AND ANALYSIS OF FACIAL ANIMATION PARAMETER PATTERNS

*Ferda Ofli, Engin Erzin, Yucel Yemez, and A. Murat Tekalp**

Multimedia, Vision and Graphics Laboratory
Koç University,
Sarıyer, Istanbul, 34450, Turkey
{fofli,eerzin,yyemez,mtekalp}@ku.edu.tr

ABSTRACT

We propose a framework for estimation and analysis of temporal facial expression patterns of a speaker. The proposed system aims to learn personalized elementary dynamic facial expression patterns for a particular speaker. We use head-and-shoulder stereo video sequences to track lip, eye, eyebrow, and eyelid motion of a speaker in 3D. MPEG-4 Facial Definition Parameters (FDPs) are used as the feature set, and temporal facial expression patterns are represented by the MPEG-4 Facial Animation Parameters (FAPs). We perform Hidden Markov Model (HMM) based unsupervised temporal segmentation of upper and lower facial expression features separately to determine recurrent elementary facial expression patterns for a particular speaker. These facial expression patterns coded by FAP sequences, which may not be tied with prespecified emotions, can be used for personalized emotion estimation and synthesis of a speaker. Experimental results are presented.

Index Terms— dynamic facial expression analysis, temporal patterns

1. INTRODUCTION

Facial expression analysis and synthesis techniques have received increasing interest in recent years. Numerous new applications can be found, for instance in the motion picture/broadcast industry for animating 3D characters and low bit-rate communications. State of the art visual speaker animation methods are capable of generating synchronized lip movements automatically from speech content; however, they lack automatic synthesis of speaker emotions and gestures from speech. Head gestures and face expressions are usually added manually by artists, which is costly and may look unrealistic.

There is an increasing interest to the facial expression analysis mostly for emotion detection. One can find thorough surveys in [1, 2, 3], and more recent works in [4, 5, 6]. [4] approaches to the problem of facial expression decomposition for recognition by using Higher-Order Singular Value Decomposition, a generalization of matrix SVD. [5] studies facial expression recognition through a Bayesian Belief Network model. [6] proposes a multiclass Support Vector Machine system of classifiers that are used to recognize either the six basic facial expressions or a set of chosen Facial Action Units. The facial expression analysis, which focuses on emotion detection, aims to identify six fundamental facial expressions: anger, disgust, fear, joy, sadness and surprise. In general, these emotional moods are represented with *static* facial expressions. However, natural facial expressions are dynamic and personal, that is they include person dependent temporal facial expression patterns. Hence, dynamic facial expression analysis addresses not specifically emotion recognition but identification of person dependent dynamic facial expression patterns. These expression patterns are valuable and can further be used for emotion identification and personalized face animation.

*This work has been supported by the European FP6 Network of Excellence SIMILAR.

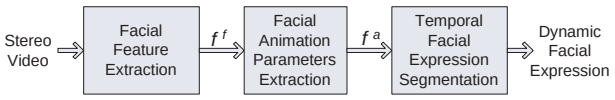

Fig. 1. Block diagram of the proposed analysis system.

In this paper, we aim to estimate 3D head motion and dynamic facial expressions of a speaker by tracking and analysis of the lip, eye, eyebrow, and head movements from stereo video sequences. The extracted expressions are represented by a set of facial animation parameters which are part of the MPEG-4 facial animation standards. A block diagram of the proposed system for facial expression patterns analysis is given in Fig. 1. There is a two-stage feature extraction module prior to analysis of the face. Feature extraction module tracks the facial expression features $\mathbf{f}^f$ in the training stereo video sequences of a speaker. Then facial animation parameters $\mathbf{f}^a$ are extracted from $\mathbf{f}^f$. At the analysis stage, facial expression feature stream is used to train a parallel HMM structure in a similar fashion explained in [7], which provides a probabilistic model for temporal recurrent facial expression patterns. The segments corresponding to these patterns are detected and labeled over the training video streams, where labels are outputted as dynamic facial expression patterns.

2. STEREO FACE ANALYSIS

Face analysis process consists of three main tasks: detecting and tracking facial feature points that correspond to Facial Definition Parameters (FDPs), extracting 3D coordinates from 2D pixel locations of FDPs, and converting the set of 3D positions of FDPs into the set of Facial Animation Parameters (FAPs) to be modeled.

2.1. Facial Feature Tracking and FAP extraction

We employed *Active Appearance Models* (AAM) approach that was introduced by Cootes, Edwards and Taylor [8, 9], as a means for modeling and tracking face components. Fig. 2 demonstrates a sequence of frames as a result of this tracking method. The output of the tracking is two sets of 2D pixel positions, $\mathbf{p}^l = \{p_1^l, p_2^l, \ldots, p_K^l\}$ for left view and $\mathbf{p}^r = \{p_1^r, p_2^r, \ldots, p_K^r\}$ for right view where K is

Fig. 2. Example image sequence that demonstrates the performance of tracking by AAMs.

the number of points tracked in each view and set to 100. $\mathbf{f}^f$ is the collection of $\mathbf{p}^l$ and $\mathbf{p}^r$, i.e., $\mathbf{f}^f = \{\mathbf{p}^l, \mathbf{p}^r\}$.

We used *MPEG-4 Facial Animation* which defines two sets of parameters: the Facial Definition Parameter (FDP) set and the Facial Animation Parameter (FAP) set [10, 11] to create the set of features for modeling facial expression patterns. These two sets provide a common framework for animating a 3D face deformable mesh model with the help of high-level and low-level facial actions, closely related to facial muscle movements.

The first set of parameters, FDPs shown in Fig. 3, is used to define feature points that are basic components in 3D face deformable meshes, represented by a 3D set of vertices. The second set of parameters, FAPs, on the other hand, consists of a collection of animation parameters that modify the positions of the FDPs. Thus, the movements of the FDPs drive the deformations to be applied to the model to animate the desired facial expressions through the transformation parameters, FAPs.

In this setting, we design a scheme to extract facial expression features, $\mathbf{f}^a$, based on MPEG-4 FAPs, using the sets of 2D pixel locations $\mathbf{f}^f$. Fig. 4 shows our 2D pixel locations to FAPs mapping scheme. We first compute the set of corresponding 3D positions $\mathbf{P} = \{P_1, P_2, \ldots, P_K\}$ of FDPs from $\mathbf{f}^f$, then calculate the set of 3D displacements $\mathbf{D} = \{D_1, D_2, \ldots, D_K\}$ of FDPs from one frame to another, and finally extract the set of FAP features $\mathbf{f}^a = \{f_1^a, f_2^a, \ldots, f_L^a\}$ from $\mathbf{D}$ where L is the number of FAPs selected to be modeled.

The first step in this process is to convert $\mathbf{f}^f$ into $\mathbf{P}$. This is simply achieved by solving the problem of stereo correspondences of the pixel locations $\mathbf{p}^l$ and $\mathbf{p}^r$. Once we perform this transformation, we have all the information for the global motion of head and face.

In order to obtain the local movements of FDPs, specifically $\mathbf{D}$, global motion of head must be subtracted from the global motions

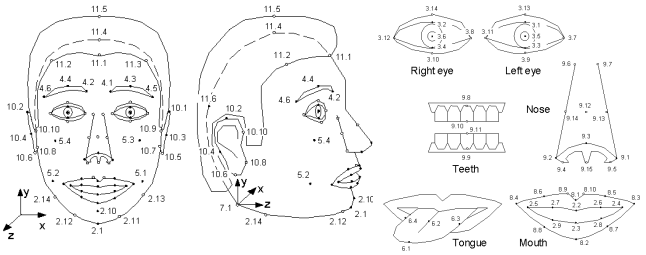

Fig. 3. The set of MPEG-4 Facial Definition Parameters. There are 84 feature points on morphological places of the neutral head model.

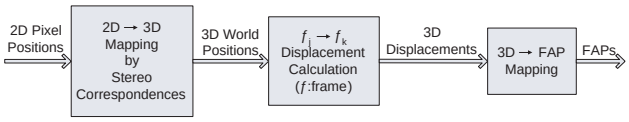

Fig. 4. Steps for converting 2D pixel values into FAPs.

of different face parts, such as lips, eyebrows, etc. Selecting a subset of FDPs that does not move locally on the face with respect to other FDPs, i.e., corners of eyes and top of nose, global motion of head is estimated with a transformation matrix $\mathbf{T}$ by least-squares fitting to two 3-D point sets that come from the reference frame and the current frame for the selected subset of FDPs [12]. Fig. 5 shows some examples of faces with respect to the initial position of face after the global head motion is undone. Then, we subtract from the current face the reference face transformed according to $\mathbf{T}$ to end up with $\mathbf{D}$. Once we have the 3D displacement of an FDP, we can select the appropriate component of its displacement vector by looking at the facial animation unit of the related FAP.

For the FAP features analysis task, we divide the face horizontally into two halves as shown in Fig. 6 and assume that face motion consists of lower face motion and upper face motion. Lower face motion contains jaw, mouth, lips and tongue which are mostly related to the lip motion due to speech utterances and emotions, whereas upper face motion contains eyes, eyebrows, eyelids which are related only to emotions. We evaluate the set of FAP features in two groups. First group is formed by the FAP features that are related to the upper face and second group is formed by the FAP features that are related to the lower face as shown in Fig. 6

2.2. Analysis

In the analysis, we try to define recurrent elementary facial expression patterns using unsupervised temporal clustering. The facial expression feature stream $\mathbf{F}$ is used to train an HMM structure $\mathbf{\Lambda}$, which capture recurrent facial expression segments ε. The parallel HMM structure $\mathbf{\Lambda}$ is used for unsupervised temporal segmentation and composed of M parallel left-to-right HMMs, $\{\lambda_1, \lambda_2, \ldots, \lambda_M\}$, where each λ_m is composed of N states, $\{s_{m,1}, s_{m,2}, \ldots, s_{m,N}\}$. The state transition matrix $\mathbf{A}_{\lambda_m}$ of each λ_m is associated with a sub-diagonal matrix of $\mathbf{A}_\Lambda$. The feature stream is a sequence of feature vectors, $\mathbf{F} = \{\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_T\}$, where $\mathbf{f}_t$ denotes the feature vector at frame t. Unsupervised temporal segmentation using HMM model $\mathbf{\Lambda}$ yields L number of segments $\varepsilon = \{\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_L\}$. The l^{th} temporal segment is associated with the following sequence of feature vectors,

$$\varepsilon_l = \{\mathbf{f}_{t_l}, \mathbf{f}_{t_l+1}, \ldots, \mathbf{f}_{t_{l+1}-1}\} \quad l = 1, 2, \ldots, L \qquad (1)$$

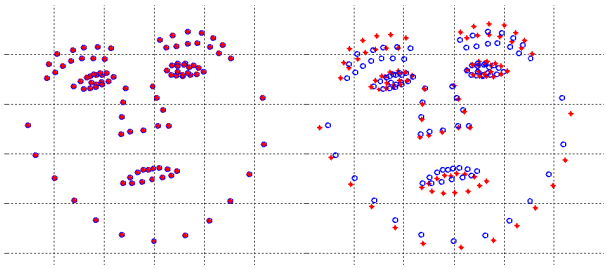

Fig. 5. Blue circles represent the reference positions of the tracked feature points, whereas Red stars represent the current positions of the tracked feature points after the global motion of the head has been removed. On the left are the the original and normalized faces drawn for the first frame, and on the right are the original and the normalized faces drawn for some other frame. Eyebrows are raised and mouth is opened from the time instant on the left to the time instant on the right. The distance between corresponding points gives the local 3D displacements of the feature points.

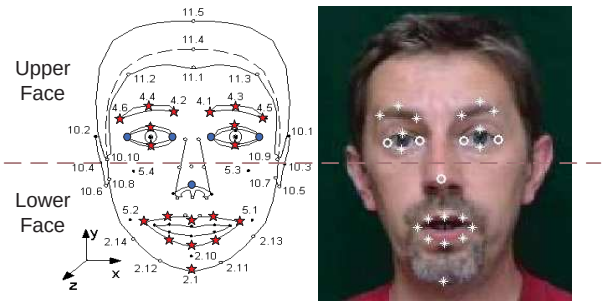

Fig. 6. The set of FDPs we are concerned at the moment. Blue circles show the set of points that we used in computing the transformation matrix, $\mathbf{T}$, and red stars show the points that we extracted FAPs for.

where $\mathbf{f}_{t_1}$ is the first feature vector $\mathbf{f}_1$ and $\mathbf{f}_{t_{L+1}-1}$ is the last feature vector $\mathbf{f}_T$.

The segmentation of the feature stream is performed using Viterbi decoding to maximize the probability of model match, which is the probability of feature sequence $\mathbf{F}$ given the trained parallel HMM $\mathbf{\Lambda}$,

$$\begin{aligned} \mathrm{P}(\mathbf{F}|\mathbf{\Lambda}) &= \max_{t_l, m_l} \prod_{l=1}^{L} \mathrm{P}(\{f_{t_l}, f_{t_l+1}, \ldots, f_{t_{l+1}-1}\}|\lambda_{m_l}) \\ &= \max_{\varepsilon_l, m_l} \prod_{l=1}^{L} \mathrm{P}(\varepsilon_l|\lambda_{m_l}) \end{aligned} \quad (2)$$

where ε_l is the l^{th} temporal segment, which is modeled by the m_l^{th} branch of the parallel HMM $\mathbf{\Lambda}$. One can show that λ_{m_l} is the best match for the feature sequence ε_l, that is,

$$m_l = \underset{m}{\operatorname{argmax}}\, \mathrm{P}(\varepsilon_l|\lambda_m) \quad (3)$$

Since, the temporal segment ε_l from frame t_l to $(t_{l+1}-1)$ is associated with segment label m_l, we define the sequence of frame labels based on this association as,

$$\ell_t = m_l \qquad \text{for } \; t = t_l, t_l+1, \ldots, t_{l+1}-1 \quad (4)$$

where ℓ_t is the label of the t^{th} frame and we have a label sequence $\ell = \{\ell_1, \ell_2, \ldots, \ell_T\}$ corresponding to the feature sequence $\mathbf{F}$. The analysis extracts the frame label sequences ℓ given the facial expression feature stream $\mathbf{F}$.

3. RESULTS

We have conducted experiments using the MVGL-MASAL database. The database includes four recordings of a single person telling stories in Turkish. Each story is approximately 7 minutes long and the total duration of the database is 27 min and 45 seconds. The audio-visual data is synchronously captured from the stereo camera and sound card. The stereo video includes only upper body with 30 frames per second whereas the audio is recorded with 16 kHz sampling rate and 16 bits per sample. The database is partitioned into two parts such that three stories are used for training of the models and one story is used for testing.

The facial expression analysis consists of unsupervised temporal segmentation of the facial expression feature stream. The parallel HMM $\mathbf{\Lambda}$ is trained with features extracted from the training video using Expectation-Maximization algorithm. The resulting HMM structure provides a probabilistic cluster model for unsupervised segmentation of facial expressions into recurring elementary patterns. We select the number of states in each branch of the face expressions HMM $\mathbf{\Lambda}$ as $N_{\mathbf{\Lambda}} = 10$, corresponding to the minimum expression pattern duration of 10 frames (1/3 sec assuming 30 video frames/sec).

We performed an iterative search for selection of the number of temporal facial expression patterns in terms of two fitness measures. The first fitness measure α is defined as the frame average of the log-probability of model match,

$$\alpha = \frac{1}{T} \log(\mathrm{P}(\mathbf{F}|\mathbf{\Lambda})) \quad (5)$$

The α measure is expected to saturate with increasing number of parallel branches in $\mathbf{\Lambda}$, since the training database is expected to contain limited number of temporal patterns. However, small variations within temporal patterns are also expected, hence the number of branches M, which saturates α measure, can be more than the actual number of temporal patterns in the training corpus. In order to make a better estimate of M, the second fitness measure β is considered as the average statistical separation between two similar temporal patterns, and it is defined as,

$$\beta = \frac{1}{T} \sum_{l=1}^{L} \log\left(\frac{\mathrm{P}(\varepsilon_l|\lambda_{m_l})}{\mathrm{P}(\varepsilon_l|\lambda_{m_l^*})}\right), \quad (6)$$

where $\lambda_{m_l^*}$ is the second best match for the temporal segment ε_l, that is,

$$m_l^* = \underset{\forall m \neq m_l}{\operatorname{argmax}}\, \mathrm{P}(\varepsilon_l|\lambda_m) \quad (7)$$

While M is increasing, the HMM branch models λ_{m_l} and $\lambda_{m_l^*}$ are expected to be similar, which decreases the β measure. Therefore, the total number of temporal patterns, M, can be selected by jointly maximizing the α and β measures. By looking at Fig. 7, we set the number of facial expression patterns $M_{\mathbf{\Lambda}}$ to 6 since the α and β measures are jointly maximized around this value.

Fig. 8 demonstrates the unsupervised clustering results for the eyebrows. We have plotted 4 FAPs, i.e., *raise_inner_eyebrow*,

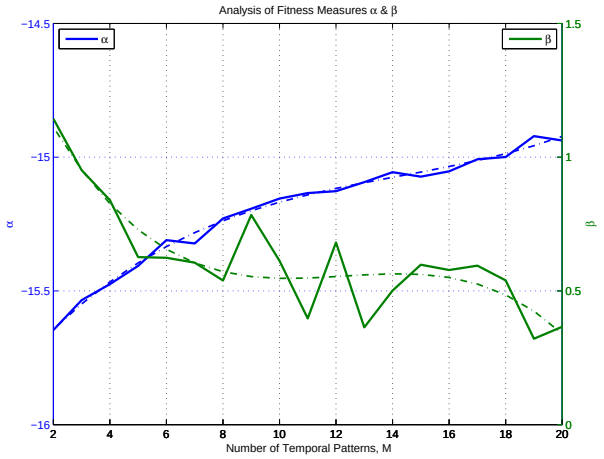

Fig. 7. The α and β fitness measures for varying number of facial expression patterns, M. The α measure, which yields the probability of model match, increases with increasing number of patterns as expected. On the other hand, the β measure, which yields statistical separation between patterns, decreases with increasing number of patterns as expected.

raise_middle_eyebrow, *raise_outer_eyebrow*, *squeeze_eyebrow*, against the state number to illustrate the dynamic pattern of the eyebrow clusterings. As can be inferred from the plots, one clustering models the raising of the eyebrows whereas the other models the reverse action of the eyebrows. Video sequences for each temporal facial expression patterns are available online [13].

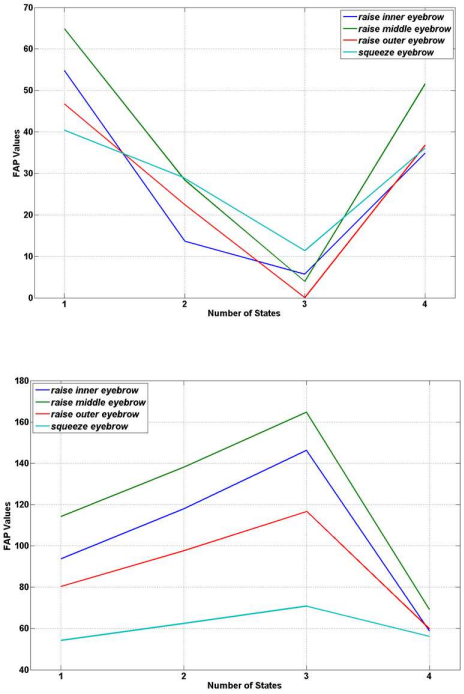

Fig. 8. Clustering results for dynamic eyebrow patterns. The upper cluster corresponds to the case where eyebrows are lowered and the lower cluster corresponds to the reverse action of eyebrows.

4. CONCLUSIONS

We have proposed a dynamic facial expression analysis system that extracts personalized recurrent elementary expression patterns for a specific user. We successfully track a user's 3D head position and orientation, as well as 3D dynamic facial expressions by using a stereo camera setup. In the analysis, we define elementary facial expression patterns for a specific speaker. Experimental results indicate that facial expressions vary from person to person, and even in time for the same person. Thus, the proposed analysis proves extremely valuable for synthesis of personalized facial expressions rather than using a generic set of facial expressions for all people.

5. REFERENCES

[1] G. Donato, M.S. Bartlett, J.C. Hager, P. Ekman, and T.J. Sejnowski, "Classifying facial actions," *IEEE Trans. on Pattern Anal. Mach. Intell.*, vol. 21, no. 10, pp. 974–989, 1999.

[2] M. Pantic and L.J.M. Rothkrantz, "Automatic analysis of facial expressions: The state of the art," *IEEE Trans. on Pattern Anal. Mach. Intell.*, vol. 22, no. 12, pp. 1424–1445, 2000.

[3] B. Fasel and J. Luettin, "Automatic facial expression analysis: A survey," *Pattern Recognition*, vol. 36, no. 1, pp. 259–275, 2003.

[4] H. Wang and N. Ahuja, "Facial expression decomposition," in *Proc. IEEE Int. Conf. on Comput. Vis.*, 2003, p. 958.

[5] D. Datcu and L.J.M. Rothkrantz, "Automatic recognition of facial expressions using bayesian belief networks," in *Proc. IEEE Int. Conf. on Systems, Man, and Cybernetics*, 2004, pp. 2209–2214.

[6] I. Kotsia and I. Pitas, "Facial expression recognition in image sequences using geometric deformation features and support vector machines," *IEEE Trans. on Image Processing*, vol. 16, no. 1, pp. 172–187, 2007.

[7] M.E. Sargin, E. Erzin, Y. Yemez, A.M. Tekalp, A.T. Erdem, C. Erdem, and M. Ozkan, "Prosody-driven head-gesture animation," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing: ICASSP 2007 (accepted to be published)*.

[8] T.F. Cootes, G.J. Edwards, and C.J. Taylor, "Active appearance models," *IEEE Trans. on Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 681–685, 2001.

[9] G.J. Edwards, C.J. Taylor, and T.F. Cootes, "Interpreting face images using active appearance models," in *Proc. IEEE Int. Conf. on Automatic Face and Gesture Recog.*, 1998, pp. 300–305.

[10] A.M. Tekalp and J. Ostermann, "Face and 2d mesh animation in mpeg-4," *Tutorial Issue On The MPEG-4 Standard, Image Communication Journal, Elsevier*, 1999.

[11] I.S. Pandzic and R. Forchheimer, *MPEG-4 facial animation : The standard, implementation and applications*, Wiley, 2002.

[12] K.S. Arun, T. S. Huang, and S.D. Blostein, "Least-squares fitting of two 3-d point sets," *IEEE Trans. on Pattern Anal. Mach. Intell.*, vol. 9, no. 5, pp. 698–700, 1987.

[13] "Video sequences for the temporal facial expression patterns," are available at `http://mvgl.ku.edu.tr/faceanalysis/`.